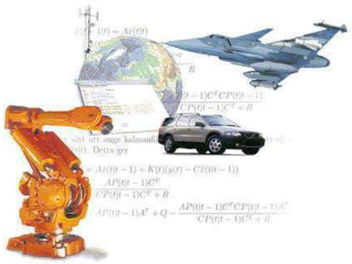
What Can Regularization Offer for Estimation of Dynamical Systems?



Lennart Ljung with Tianshi Chen

Reglerteknik, ISY, Linköpings Universitet

Outline

- Preamble: The classic, conventional System Identification Setup
- Bias – Variance, Model Size Selection
- Regularization
  - Well tuned bias–variance trade-off
  - Filling out missing information in data

State-of-the-Art System Identification

Models:

Model Structure: $\mathcal{M}$. Parameters: θ. Model: $\mathcal{M}(\theta)$.
Observed input–output (u, y) data up to time t: Z^t
Model described by predictor: $\mathcal{M}(\theta) : \hat{y}(t|\theta) = g(t, \theta, Z^{t-1})$.

Estimation:

$-$log likelihood function $\varepsilon(t, \theta) = y(t) - \hat{y}(t|\theta)$
$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t, \theta)|^2$
"Prediction Error Fit"
$\hat{\theta}_N = \arg\min V_N(\theta)$

Model Structure (size) determination, AIC, BIC:

$\mathcal{M}(\hat{\theta}_N) = \arg\min_{\mathcal{M},\theta}[\log V_N(\theta) + g(N)\text{dim}\theta]$
$g(N) = 2$ or $\log N$

Comment on Model Structure Selection

The model fit as measured by $\sum_{t=1}^{N} |y(t) - \hat{y}(t|\theta)|^2$ for a certain set of data will always improve as the model structure becomes larger (more parameters). The parameters will start adjusting also to the actual noise effects in the data ["Overfit"]
There are two ways of counteracting this effect:

- Compute the model on one set of (estimation) data and evaluate the fit on another (validation) data set. [Cross-Validation]
- Add a penalty term to the criterion which balances the overfit:

$$\mathcal{M}(\hat{\theta}_N) = \arg\min_{\mathcal{M},\theta}[\log V_N(\theta) + g(N)\text{dim}\theta]$$

$$AIC : g(N) = 2, \quad BIC : g(N) = \log(N)$$

AIC: Akaike's Information Criterion. BIC: Bayesian Information Criterion [= MDL: Minimum Description Length]

Model Estimate Properties

As the number of data, N, tends to infinity

- $\hat{\theta}_N \to \theta^* \sim \arg\min_\theta E|\varepsilon(t,\theta)|^2$ the best possible predictor in $\mathcal{M}$
- If $\mathcal{M}$ contains a true description of the system
  - Cov $\hat{\theta}_N = \frac{\lambda}{N}[E\psi(t)\psi^T(t)]^{-1}$ $[\psi(t) = \frac{d}{d\theta}\hat{y}(t|\theta), \lambda$: noise level]...
  - ... is the Cramér-Rao lower bound for any (unbiased) estimator.

E: Expectation. These are very nice optimal properties:

- The model structure is large enough: The ML/PEM estimated model is (asymptotically) the best possible one. Has smallest possible variance (Cramér- Rao)
- The model structure is not large enough: The ML/PEM estimate converges to the best possible approximation of the system. ”The estimate has smallest possible asymptotic bias.”

Linear Models

General Description

$y(t) = G(q,\theta)u(t) + H(q,\theta)e(t)$, q : shift op. e : white noise

$$G(q,\theta)u(t) = \sum_{k=1}^{\infty} g_k u(t-k), \quad H(q,\theta)e(t) = 1 + \sum_{k=1}^{\infty} h_k e(t-k)$$

Predictor

$$\hat{y}(t|\theta) = G(q,\theta)u(t) + [I - H^{-1}(q,\theta)][y(t) - G(q,\theta)u(t)]$$

Asymptotics: $[\Phi_u, \Phi_v$: Spectra of input and additive noise $v = He$.]

$$\hat{\theta}_N \to \theta^* = \arg\min_\theta \int_{-\pi}^{\pi} |G(e^{i\omega},\theta) - G_0(e^{i\omega})|^2 \frac{\Phi_u(\omega)}{|H(e^{i\omega},\theta)|^2} d\omega$$

$$\text{Cov} G(e^{i\omega}, \hat{\theta}_N) \sim \frac{n}{N}\frac{\Phi_v(\omega)}{\Phi_u(\omega)} \text{ as } n, N \to \infty \quad n : \text{model order}$$

Common Black-Box Parameterizations:

BJ (Box-Jenkins)

$$G(q,\theta) = \frac{B(q)}{F(q)}; \quad H(q,\theta) = \frac{C(q)}{D(q)}$$
$$B(q) = b_1 q^{-1} + b_2 q^{-2} + \ldots b_{nb} q^{-nb}$$
$$F(q) = 1 + f_1 q^{-1} + \ldots + f_{nf} q^{nf}$$
$$\theta = [b_1, b_2, \ldots, f_{nf}]$$

ARX:

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \text{ or}$$
$$A(q)y(t) = B(q)u(t) + e(t) \text{ or}$$
$$y(t) + a_1 y(t-1) + \ldots + a_{na} y(t-na)$$
$$= b_1 u(t-1) + \ldots + b_{nb} u(t-nb)$$

Common Black and Grey Parameterizations

State-Space with Possibly Physically Parameterized Matrices

$$x(t+1) = A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t)$$
$$y(t) = C(\theta)x(t) + e(t)$$

Corresponds to

$$G(q,\theta) = C(\theta)(qI - A(\theta))^{-1}B(\theta).$$
$$H(q,\theta) = C(\theta)(qI - A(\theta))^{-1}K(\theta) + I$$

Continuous Time (CT) Models

Physical Model with unknown parameters

$$\dot{x}(t) = \mathcal{F}(\theta)x(t) + \mathcal{G}(\theta)u(t) + w(t)$$
$$y(t) = C(\theta)x(t) + D(\theta)u(t) + v(t)$$

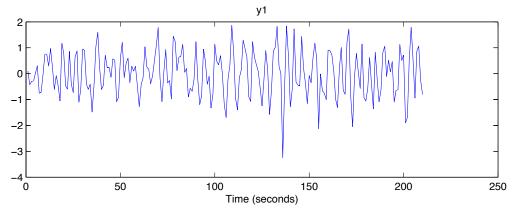
Sample it (with correct Input Intersample Behaviour):

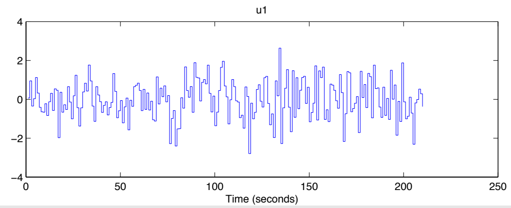
$$x(t+1) = A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t)$$
$$y(t) = C(\theta)x(t) + e(t)$$

Now apply the discrete time formalism to this model, which is parameterized in terms of the CT parameters θ

An Example

Equipped with these tools, let us now test some data (selected but not untypical). The example uses complex dynamics and few (210) data, so this is a case where asymptotic properties are not important.

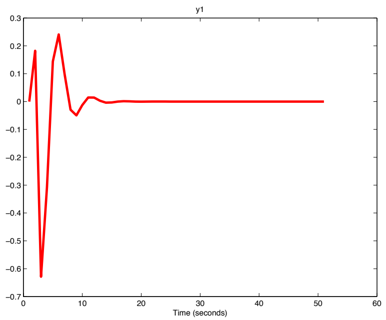




Estimate a Model: State-of-the-Art

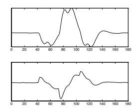
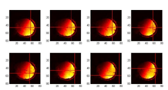
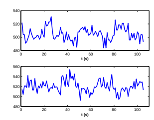
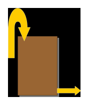
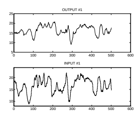
We will try the state-of-the art approach: Estimate SS models of different orders. Determine the order by the AIC criterion.

```
 for k=1:30
   m{k}= ssest(z,k);
end
(dum,n) = min(aic(:});
mss = m{n};
impulse(mss)
```



Is this a good model? Preview: This IR has a fit of **79.42%**
But, we can do better! Another choice of model order gives a fit of **82.95 %** . I will also show an estimate with **a 83.56%** fit.

Status of the State-of-the-Art Framework

- Well established statistical theory
- Optimal asymptotic properties
- Efficient software
- Many applications in very diverse areas. Some examples:
  - Aircraft Dynamics:  
  - Brain Activity (fMRI):  
  - Pulp Buffer Vessel:   ;

Time-out

This is a bright and rosy picture. Any issues and problems?

- Convexity Issues: For most model structures the criterion function $V_N(\theta) = \sum_{t=1}^{N} |y(t) - \hat{y}(t|\theta)|^2$ is non-convex and multi-modal (several local minima). *Evolutionary Minimization Algorithms* could be applied, but no major successes for identification problems have been reported. We typically have to resort to good initial estimates.
- Small data sizes – complex systems: Need well tuned bias–variance trade–off. Model selection rules are a bit shaky in this case. [Recall: "We can do better."]

Bias – Variance Trade Off

Any estimated model is incorrect. The errors have two sources:

- Bias: The model structure is not flexible enough to contain a correct description of the system.
- Variance: The disturbances on the measurements affect the model estimate, and cause variations when the experiment is repeated, even with the same input.

Mean Square Error (MSE) = $|\text{Bias}|^2$ + Variance.
When model flexibility ↑,Bias ↓ and Variance ↑.
To minimize MSE is a good trade-off in flexibility.
In state-of-the-art Identification, this flexibility trade-off is governed primarily by model order. May need a more powerful tuning instrument for bias–variance trade-off.

Linear Black-Box Models: Fundamental Role of ARX

ARX can Approximate Any Linear System

Arbitrary Linear System: $y(t) = G_0(q)u(t) + H_0(q)e(t)$

ARX model order n, m : $A_n(q)y(t) = B_m(q)u(t) + e(t)$

as $N >> n, m \to \infty$

$[\hat{A}_n(q)]^{-1}\hat{B}_m(q) \to G_0(q), \ [\hat{A}_n(q)]^{-1} \to H_0(q)$

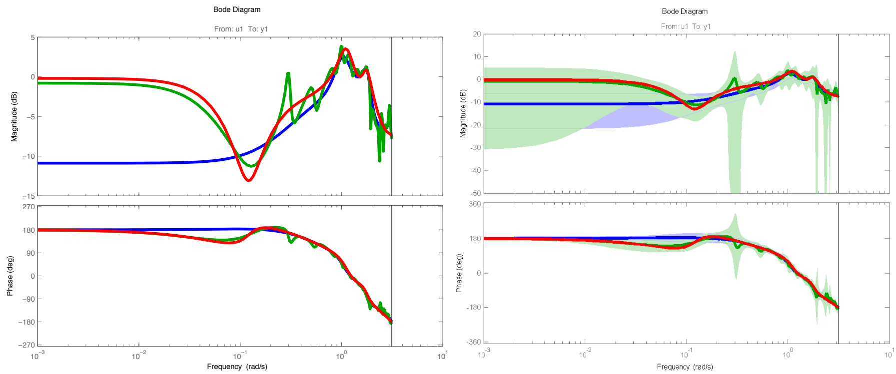
The ARX-model Is a Linear Regression

Note that the ARX-model is estimated as a linear regression $Y = \Phi\theta + E$, (Φ containing lagged y, u and θ containing a, b)
A convex estimation problem.

Virtually all methods to find a linear intial estimate for the non-convex minimization of the ML criterion are based on an ARX-model of some kind.

Test on Our Data

Estimate ARX-model of order 10 and 30: Bode plots of models together with true system:



Order 10. Order 30. True. The high order model picks up the true curves better, but seem more "shaky". Look at Uncertainty regions!

How to Curb Variance?

The ARX approximation property is valuable, but high orders come with high variance.
Can we curb the flexibility that causes high variance other than by lower order? Regularization

Model Structures and Regularization

Curb the Model's Flexibility!

$$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta-\theta^*)^T R(\theta-\theta^*)$$

Bayesian View

The regularized criterion

$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta-\theta^*)^T R(\theta-\theta^*)$

Bayesian interpretation: Suppose θ is a random vector which *a priori*

$$\theta \in N(\theta^*, \Pi), \quad f(\theta) = \frac{1}{\sqrt{(2\pi)^d \det(\Pi)}} e^{-(\theta-\theta^*)^T \Pi^{-1}(\theta-\theta^*)/2}$$

Bayes rule gives *posterior dist* (Y denoting all measured y-signals)

$$P(\theta|Y) = \frac{P(\theta,Y)}{P(Y)} = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

Apart from the normalization, and other θ-independent terms, twice the negative logarithm of $P(\theta|Y)$ is $V_N(\theta)$ with $\lambda R = \Pi^{-1}$
That means that with the regularized estimate $\hat{\theta}_N = \arg\min V_N(\theta)$ is the *Maximum A Posteriori* (MAP) Estimate.

Tuning the Regularization

The regularized criterion

$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta-\theta^*)^T R(\theta-\theta^*), \quad \lambda R = \Pi^{-1}$

- $R = I, \theta^* = 0$, tune λ: Ridge regression
- Cross Validation
- Use ML for marginal distribution ("Empirical Bayes"): Parameterize $\theta^*(\alpha)$, $\Pi(\alpha)$ with *hyper-parameters* α. Form

$$P(Y|\alpha) = \int P(Y|\theta,\alpha)P(\theta|\alpha)d\theta$$

$$\hat{\alpha} = \arg\max P(Y|\alpha)$$

First factor essentially the likelihood function for θ and second factor essentially the prior. The integration is simple for a linear regression model, see next few slides.

Linear Regression – Regularization

The regularized criterion

$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta-\theta^*)^T R(\theta-\theta^*), \quad \lambda R = \Pi^{-1}$

Regularization for a linear regression ($\theta^* = 0$)

$$Y = \Phi\theta + E$$
$$\hat{\theta}_N = \arg\min |Y - \Phi\theta|^2 + \theta^T \Pi^{-1}\theta$$

Π is the Regularization Matrix. MSE:

$$\mathcal{E}[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T] = (R_N + \Pi^{-1})^{-1} \times$$
$$(R_N + \Pi^{-1}\theta_0\theta_0^T\Pi^{-1})(R_N + \Pi^{-1})^{-1} \qquad R_N = \Phi\Phi^T,\ \theta_0 = \text{true par}$$

Minimized by $\Pi = \theta_0\theta_0^T$: MSE = $(R_N + \Pi^{-1})^{-1}$ How to select Π?

Marginal Likelihood for Regularized Linear Regression

Recall Empirical Bayes (EB): Parameterize $\theta^*(\alpha),\ \Pi(\alpha)$. Form

$$P(Y|\alpha) = \int P(Y|\theta,\alpha)P(\theta|\alpha)d\theta$$
$$\hat{\alpha} = \arg\max P(Y|\alpha)$$

In the linear regression case

$Y = \Phi\theta + E, \quad \theta \in N(0,\Pi(\alpha)),\ E \in N(0,I),\ \Phi$ deterministic

$$Y = \Phi\theta + E \in N(0, Z(\alpha)), \quad Z(\alpha) = \Phi\Pi(\alpha)\Phi^T + I$$

$P(Y|\alpha)$ immediate

ML estimate of α : $\quad \hat{\alpha} = \arg\min Y^T Z(\alpha)^{-1} Y + \log\det Z(\alpha)$

ARX Model Priors

When estimating an ARX-model, we can think of the predictor

$$\hat{y}(t|\theta) = (1 - A(q))y(t) + B(q)u(t)$$

as made up of two impulse responses, A and B. The vector θ should thus mimic two impulse responses, both typically exponentially decaying and smooth.We can thus have a reasonable prior for θ:

$$P(\alpha_1,\alpha_2) = \begin{bmatrix} P^A(\alpha_1) & 0 \\ 0 & P^B(\alpha_2) \end{bmatrix} \qquad \text{Block Diagonal } A\&B$$

where the hyperparameters α describe decay and smoothness of the impulse responses. Typical choice:

TC kernel

$E|b_k|^2 = C\lambda^k$, corr$(b_k, b_{k+1}) = \sqrt{\lambda}$
$P^B_{k,\ell} = C\min(\lambda^k, \lambda^\ell); \quad \alpha = [C, \lambda]$

Relevant References

This classical regularization framework/Bayesian tuning framework was suggested for the estimation of linear systems impulse responses by
Pillonetto, De Nicolao and Chiuso
in 2010/2011 (Automatica/IEEE AC) using a function learning perspective.

The current classical regularization interpretation was made by
Chen, Ohlsson and Ljung
in 2011/2012 (IFAC WC/Automatica).

Software Issues

The MATLAB system Identification Toolbox, ver R2013b (released August 2013) now supports quadratic regularization for all linear and non-linear model estimation.

The regularized criterion

$$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta - \theta^*)^T R(\theta - \theta^*),$$

is supported by a field `Regularization` in all the estimationOptions (`arxOptions, ssestOptions, procestOptions`) etc.:

```
opt.Regularization.Lambda
opt.Regularization.R
opt.Regularization.Nominal
```
(θ^*)

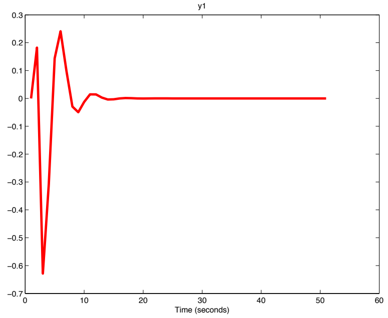
ARX-regularization tuning:

```
[L,R]=arxRegul(data,[na,nb,nk],Kernel)
```


Our Test Data: State-of-the-Art

Recall: The state-of-the art approach: Estimate SS models of different orders. Determine the order by the AIC criterion.

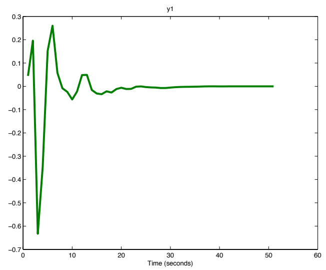
```
 for k=1:30
   m{k}= ssest(z,k);
end
(dum,n) = min(aic(:));
mss = m{n};
impulse(mss)
```



Estimate a Model: Regularized ARX

Now, let us try an ARX model with na=5, nb=60. Estimate a regularization matrix with the 'TC' kernel (2 parameters, C, λ each for the A and B parts):

```
 aopt = arxOptions;
(L,R) = arxRegul(z,[5 60 0],'TC');
aopt.Regularization.R = R;
aopt.Regularization.Lambda = L;
mr = arx(z,[5 60 0],aopt);
impulse(mr)
```

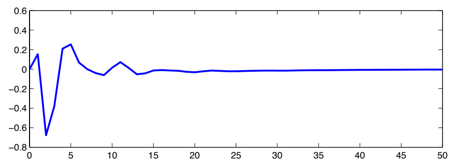


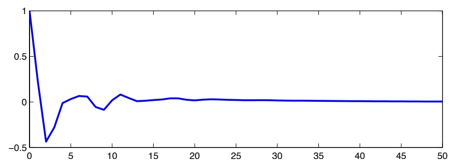
The Oracle

The examined data were obtained from a randomly generated model of order 30:

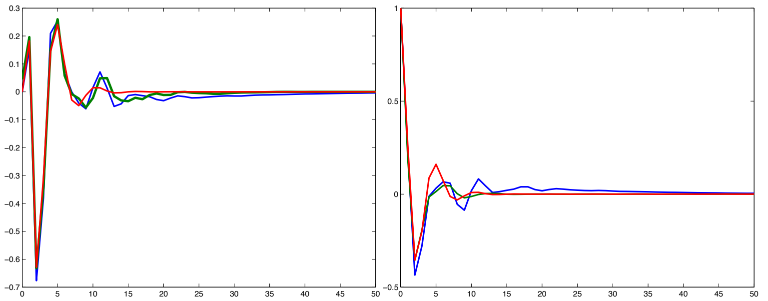
$$y(t) = G_0(q)u(t) + H_0(q)e(t)$$

The input is Gaussian white noise with variance 1, and e is white noise with variance 0.1. The impulse responses of G and H are shown at the right.





How Well Did Our Models mss And mr Do?



G : fit: **mss: 79.42% mr: 83.55%** H: fit **mss: 77.05%**, **mr: 81.59%**

Surprise ?

ML beaten by an "outsider algorithm"!: That is a surprise!
There is a certain randomness in these data, but Monte-Carlo simulations substantiate the observed conclusion.

Even though ML is known to have the quoted optimal properties for best bias and variance, the observation is still not a contradiction.

Recall: Mean Square Error (MSE) = $|\text{Bias}|^2$ + Variance.

ML: Bias $\approx 0 \Rightarrow$: MSE = Variance = CR Lower bound for unbiased estimators

But with some bias, Variance could be clearly smaller then CRB

Recall for Lin Reg: CRB = $(\Phi\Phi^T)^{-1} > (\Phi\Phi^T + P^{-1})^{-1}$ = MSE for best regularized estimated. More pronounced for short data

Objections?

Recall: mss fit 79.42%, mr fit 83.55 %

- We were just unlucky to pick order 3 (AIC). Other model selection criteria would have given better results.
  - If we ask the oracle what is the best possible state-space order for ML estimated model, the answer is **order 12 for G with a fit 82.95 %** and **order 3 for H with a fit 77.04%** So the regularized ARX -model gives better fit to both G and H than is at all possible for ML estimated state-space models [for these data].
- The R-ARX model is of order 60, and it is unfair to compare it with SS models of low order.
  - Try `mred = balred(mr,7)` to create a 7th order SS-model. It still has a G-fit of **83.56%** and outperforms the oracle-selected ML SS models.

Discussion

- In this case Regularized ARX gave a much better and more flexible bias–variance trade off through the continuously adjustable hyperparameters in the regularization matrix — Compared to the state-of-the art bias–variance trade off in terms of discrete model orders.
- Can we forget about `ssest` and move over to regularized ARX?
  - No, recall that the studied situation had quite few data, and the good trade-off is reached for rather large bias, not favouring ML.
  - But one should be equipped with regularized ARX in one's toolbox
- Regularized ARX (possible followed by `balred`) can be seen as a convexification of the state-of-the art SS model estimation techniques.
  NB: Tuning of hyperparameters normally non-convex

FIR Impulse Response Estimation

The system:

$$G(z) = \frac{0.02008 + 0.04017z^{-1} + 0.02008z^{-2}}{1 - 1.561z^{-1} + 0.6414z^{-2}} \tag{1}$$

Impulse response:

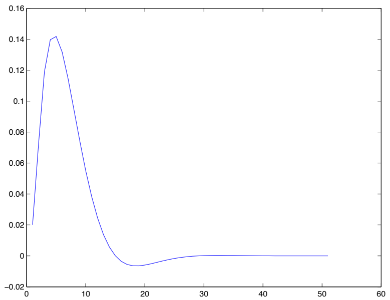


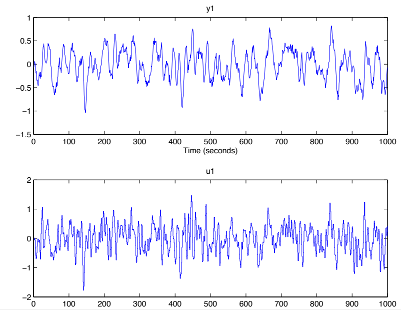
Figure : The true impulse response.

The Model and the Data

The model:

$$y(t) = \sum_{k=0}^{nb} g(k)u(t-k) \tag{2}$$

The data: (Input low pass filtered white noise, white measurement noise: SNR: 400)



Best FIR model (nb=13)

```
arx(z,[0 13 0])
```

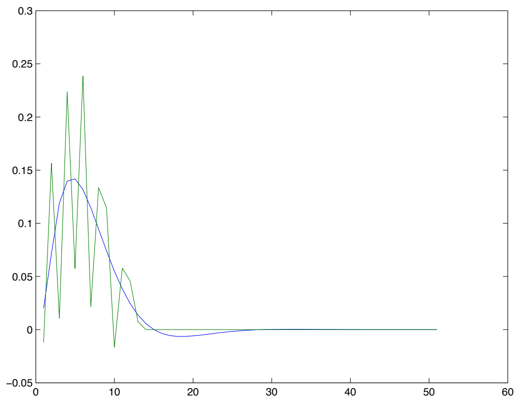


Figure : The true impulse response together with the estimate for order $nb = 13$.

Ridge Regression (nb=50)

```
aopt=arxOptions; aopt.Regularization.Lambda=1;
m50r=arx(z,[0 50 0],aopt);
```

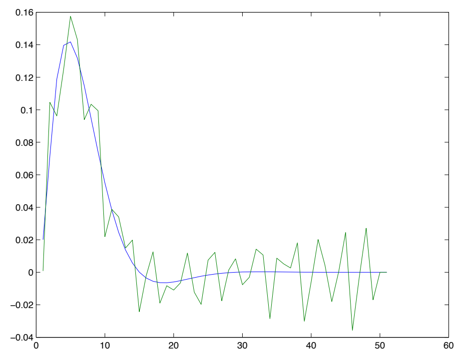


Figure : The true impulse response together with the ridge-regularized estimate for order $nb = 50$.

Tuned Regression

[L,R]=arxRegul(z,[0 50 0],'TC'); aopt.Regularization.Lambda=L; aopt.Regularization.R=R; mrtc=arx(z,[0 50 0],aopt);

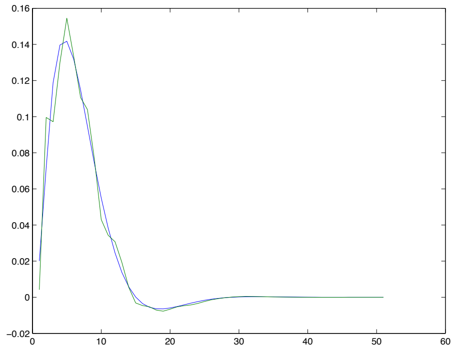
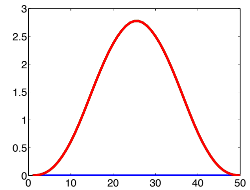


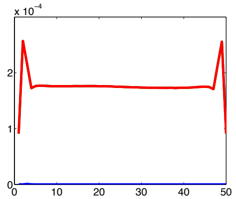
Figure : The true impulse response together with the tuned regularized estimate for order $nb = 50$.

Bias-Variance Trade–off for FIR-50 Model

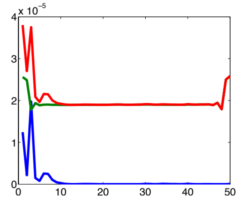
MSE = BIAS^2 + Variance (function of lag for IR)



The unregularized estimate. The Bias $\equiv 0$. MSE = Variance = The CR Lower Bound



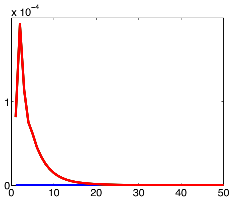
The Ridge Regression Estimate with $\lambda = 1$. Variance has decreased, bias is still negligable.



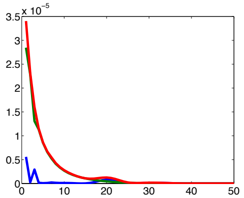
The Ridge Regression Estimate with $\lambda = 10$. The bias has increased.

Bias-Variance Trade–off for FIR-50 Model

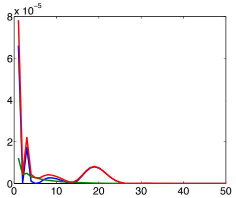
MSE = BIAS^2 + Variance (function of lag for IR)



The regularized estimate (EB)



The regularized (EB) estimate with λ increased 10 times.



The regularized (EB) estimate with λ increased 50 times.

Discussion

- In this case the main reason for the poor conventional estimates was the poor input excitation at high frequencies
- The simple prior "smooth decaying IR" (with the numerical details being estimated) was sufficient to fill out this lacking information.

Conclusions

- The State-of-the art system identification relies upon a solid statistical ground, with (ML-like) parameter estimation in chosen model structures.
- The bias-variance trade-off in terms of model order could be unsatisfactory, esp. for smaller data sets.
- Regularization is well known in statistics, but has not been used so much in system identification.
- Regularized ARX-models offer a fined tuned choice for efficient bias–variance trade-off and form a viable convex alternative to state-of-the-art ML techniques for linear black-box models.
- Regularization also offers important complements for inadequate information contents in data.

References and Acknowledgments

- The regularization results were based on and inspired by: T. Chen, H. Ohlsson and L. Ljung: On the estimation of transfer functions, regularization and Gaussian Processes – Revisited. *Automatica*, Aug 2012.
- Funded by the ERC advanced grant LEARN

European Research Council
Established by the European Commission